Response to referee 3:
We thank the reviewer for his/her comments. Below is our response in blue:

In the course of responding to the reviewers, two main changes occurred: (1) a trend discussion was added and (2) we emphasize the lack of MLS penetration below the upper troposphere throughout the paper.

Trend discussion

Before the conclusions the following paragraphs will be added:

As an example, Figure 8 shows the H_2O and O_3 trends in the tropics computed using monthly zonal mean deseasonalized anomalies of the raw model fields, as well as using all the available satellite-sampled measurement locations and only those passing the screening criteria in the tropics. As shown, when all available measurement locations are used, the MIPAS and MLS sampling allows accurate derivation of trends, with values matching those calculated from the raw model fields almost exactly. However, when only those measurements passing the screening criteria are used, both instruments have limitations: MIPAS trends are impacted because of the large percentage of measurements screened out below 100 hPa, which introduces non-negligible artifacts (up to 80% change for H_2O and up to 20% change for O_3); MLS trends are impacted because of the reduced vertical resolution, which limits its usefulness to the upper troposphere and above. Note that the impact of quality screening on MIPAS trends can be mitigated by using a regression model similar to the ones used by Bodecker et al. (2013) and Damadeo et al. (2014). These models have been shown to mitigate the effects of the non-uniform temporal, spatial and diurnal sampling of solar occultation satellite measurements. Furthermore, MIPAS trend analysis can be restricted to regions less affected by deep convection (for example, the mid tropical Pacific) to minimize the quality screening effects.

The estimated number of years required to definitively detect these trends is also shown in Figure 8. These estimates were computed assuming a trend model similar to the one described by Tiao et al. (1990), Weatherhead et al. (1998), and Millán et al. (2016), with a seasonal mean component represented by the monthly climatological means. As shown, with the MIPAS screened fields additional years are required for robust trend detection (up to ~150 years for H_2O and up to ~40 years for O_3 versus 50 years and 20 years, respectively, when all available measurements are used).

Similar analyses were performed for other latitude bands. Although the magnitude of the trends derived when using the MIPAS screened measurement locations was also impacted in these cases, no significant difference was found in the number of years required to detect such trends. In addition, no significant artifacts were found for HNO_3, CO or temperature for either the trend magnitude or the number of years required to detect such trends. Note that, when using real data, the effect of instrument noise upon trends will be negligible due to the vast number of MIPAS or MLS measurements associated with each monthly latitude bin. Drifts and long-term stability issues on these datasets [i.e., Eckert et al., 2014; Hubert et al., 2016; Hurst et al., 2016] will have to be corrected.

In the conclusion section, the trend discussion will be changed to:
These biases affect trends derived from these measurements using a simple regression upon monthly zonal mean data substantially affected by clouds. Further, the number of years required to detect such trends may increase due to the extra noise added to the time series by screening out measurements.

The following figure will be added (as figure 8 of the revised paper):

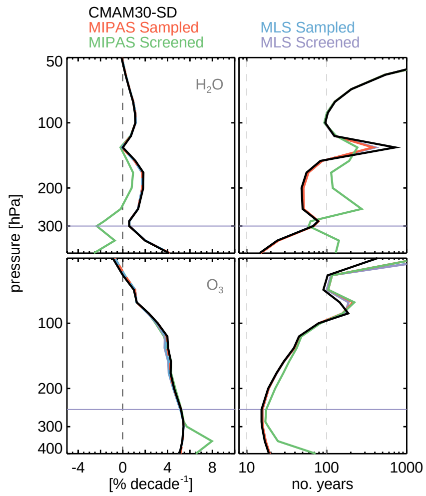


Figure 8: (left) H_2O and O_3 trends computed based on monthly zonal mean deseasonalized anomalies for the tropics (20S to 20N) using the raw model fields, all the available satellite measurements (MIPAS or MLS sampled) and only those measurements passing the screening criteria (MIPAS or MLS screened). Note that for O_3, we only use data starting from 2000 to capture the expected period of O_3 recovery. A purple line indicates the bottom (largest pressure) of the recommended range of the MLS retrievals. (right) Number of years required to detect such trends.

References:
Bodecker et al 2013: 10.5194/essd-5-31-2013
Damadeo et al 2014: 10.5194/acp-14-13455-2014
Tiao et al. (1990): 10.1029/JD095iD12p20507
Weatherhead et al. (1998): 10.1029/98JD00995
Millán et al (2016): 10.5194/acp-16-11521-2016

Lack of MLS penetration

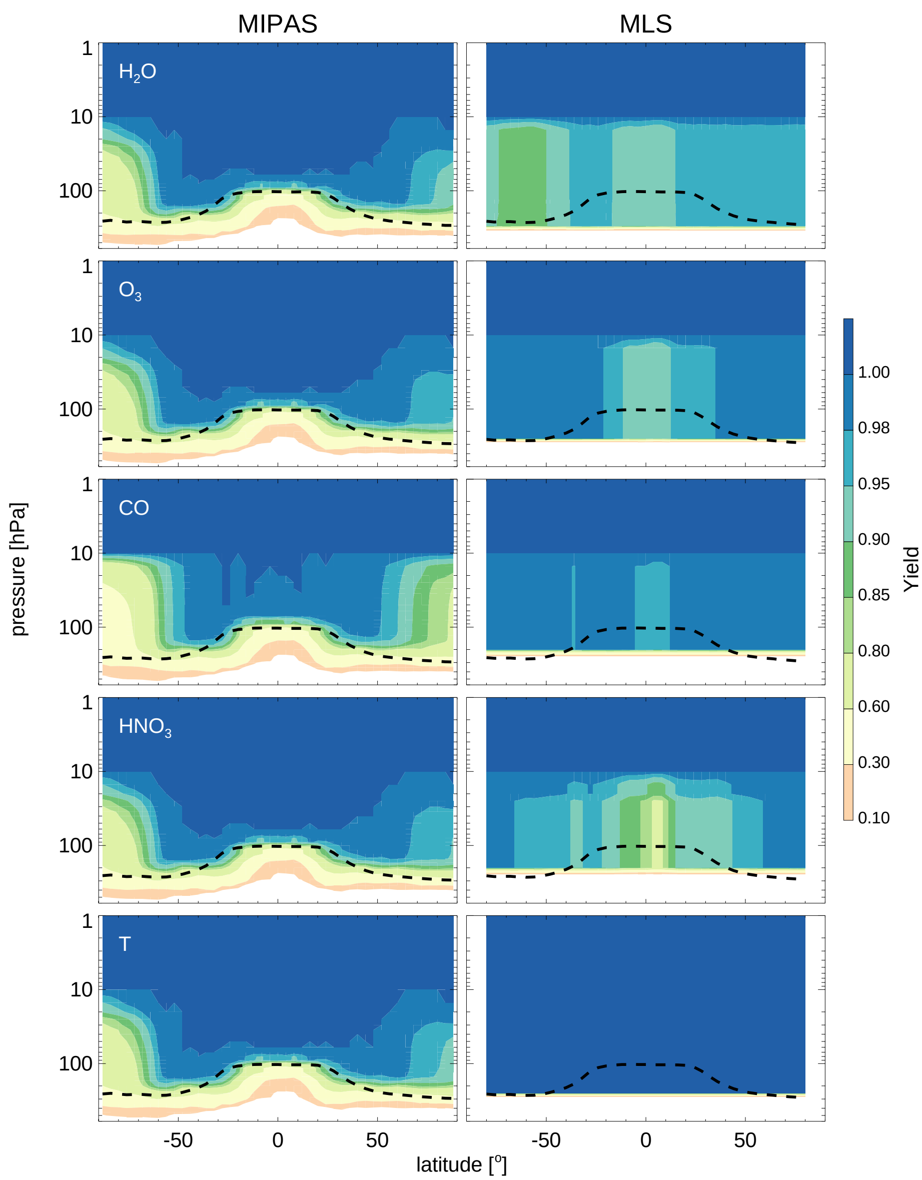
To emphasize more the MLS caveat, the sentence in P1L16 (in the abstract) will be changed to: In contrast, MLS data quality screening removes sufficiently few points that no additional bias is introduced, although its penetration is limited to the upper troposphere while MIPAS may cover well into the mid troposphere in cloud-free scenarios.

In P5 L8 this sentence will be changed: In contrast, in general MLS yield values are better than 90%, although the measurements do not extend below the upper troposphere.

In a similar manner P7 L18 will be changed to: However, continuum absorption in the microwave suppresses signals from the mid and lower troposphere in a limb viewing geometry, limiting the MLS vertical range to the upper troposphere and above while MIPAS may cover well into the mid troposphere in cloud free scenes.

And in the new paragraph about trends we included: However, when only those measurements passing the screening criteria are used, both instruments have limitations: MIPAS trends are impacted because of the large percentage of measurements screened out below 100 hPa, which introduces non-negligible artifacts (up to 80% change for H_2O and up to 20% change for O_3); MLS trends are impacted because of the reduced vertical resolution, which limits its usefulness to the upper troposphere and above.

Furthermore, we noticed that the Figures were not displaying the correct MLS pressure cut off. The revised versions showcase much better the lack of MLS penetration (as an example, the updated Figure 2 is shown below). Also, we superimpose a mean thermal tropopause derived from MERRA2 to Figure 2, 3, and 4.



Updated Figure 2

Reviewer comments

The paper by Millan et al. describes effects of sampling and quality screening for two limb sounding instruments (MLS and MIPAS). Based on CCM data the analysis shows significant biases in the zonally averaging of various trace gases and temperature distribution, especially for a MIPAS-like instruments with its cloud-induced quality screening. Biases in the zonal means as well as for trend analyses in the upper troposphere can be quite large, which is an important fact and needs consideration in any trend analysis.

General comments:

There are two major concerns regarding the manuscript from Millan et al.:

a) The study highlights a very specific caveat of sampling and screening issues of space borne instruments. The whole approach is very technical, investigating a very obvious problem of the IR limb measurements due to cloud screening, which can have a severe effect on the results especially if you like to analyse trends in the very cloudy upper tropical troposphere. Although this effect is of interest for ACP community, these kind of analysis is not really suited for ACP. The whole manuscript looks more like a technical study or a technical note (for example as prepared for NASA or ESA studies for future satellite sensors) and is not sufficient for a full publication in a journal like ACP. A more in-depth study of the MW/IR limb caveats on trend analyses cloud be publishable in a more technical and instrument orientated journal like Atmospheric Measurement Techniques.

As the referee notes, determination of trends is a critical issue of great current concern. While the effects of clouds on IR measurements are indeed well known, the impact of removing cloud-affected profiles on biases in zonal means and the trends estimated from them has not been quantified before. Hence, we believe that this paper is appropriate for ACP because its results directly impact the broad UTLS community.

b) To my mind, this paper does not contribute a substantial add-on to a former study of the authors. The paper by Millan et al. (2016) in ACP describes most of the general effects of sampling pattern, which can be quite similar effects to the quality screening, in the context of a much more detailed study on atmospheric trend analyses. My strong impression is that the actual manuscript misses new concepts and methods and just repeats parts of the Millan et al. (2016) analyses with another combination of satellite instruments.

The new addition is the quantification of the screening bias *on top of* the sampling bias. Previous studies (Toohey et al., 2013; Sofieva et al., 2014; Millán et al., 2016) focused on the sampling bias while disregarding the effects of data quality screening. As shown in Figure 3 and 4 of the paper, the effects of the screening biases are not insignificant and need to be known and quantified. A different combination of satellite instruments was essential to this study because the idea was to take data sets with comparable sampling biases and show how they diverge when screening biases are taken into account.

P2 L17 will be expanded: However, none of these studies have quantified the additional biases introduced through quality screening of the measurements, that is, this study isolates and quantifies another source of uncertainty in averaged data.

Major comments:

1) P4, line 29: Although the screening is a fundamental issue of the study, there are no details presented for MLS (but some for MIPAS). Please, at least summarize the main topics of Livesey et al. (2006). In addition, more details on the cloud screening of MIPAS would be helpful as well (see next comment).

We will add: The screening procedure applied to MLS data follows the guidelines detailed by Livesey et al. (2017): We only use data within the specified pressure ranges; we neglect profile points for which the precision is negative (which indicates that the retrievals are influenced by the a priori); we avoid profiles for which the "Status" field is an odd number (which indicates operational abnormalities or problems with the retrievals); and we only use profiles for which the "Quality" and "Convergence" fields are within the specified thresholds. The "Quality" field describes the degree to which the measured radiances have been fitted by the retrieval algorithm and the "Convergence" field is a ratio of the fit achieved at the end of the retrieval process to the value predicted at the previous step.

2) P5, line 1: Please note somewhere here that the MIPAS cloud-screening applied in the IMK product is a very conservative approach with respect to cloud effects in the measured spectra. This might have effects on the presented analysis. Especially, in the tropical upper troposphere high water vapour abundance will result in false-positive cloud detections (e.g. Spang et al. 2012) in the IMK product and consequently an artificial overestimation in the biases, where other MIPAS level 2 processors might do a better job. This objective should be addressed in the manuscript.

The following sentence will be added in P5 line 28: We note that the cloud screening procedure of the IMK/IAA algorithm is conservative with respect to those of other MIPAS processors [Spang et al., 2012]. On the face of it, it appears that this causes an unnecessarily large sampling bias which could be avoided by using a less restrictive cloud screening threshold value. The purpose of a conservative cloud screening procedure is to guarantee that the measurements passing the cloud screening are indeed unaffected by any cloud signal in the spectra. Cloud signals can lead to systematic retrieval errors, which are correlated to the state of the atmosphere, hence, the sampling biases would merely be replaced by retrieval biases. This, we think, is worse, because then both the parent data and the zonal averages would be affected, while with a conservative screening only the averages are affected but not the parent data which survive the screening.

3) The disadvantages of the cloud screening in the tropical UT for the IR measurements are highlighted frequently in the manuscript and are presented by different types of analyses, but with only very limited further information content. In contrast, the caveat of MW sounders to retrieve trace gases below ~300 hPa at any latitude is only briefly mentioned. Studies for future ESA missions, like the PREMIER report for mission selection (http://projects.knmi.nl/capacity/PREMIER/SP1324-3_PREMIERr.pdf), show nicely the excellent synergy by a combination of IR and MW (s. Fig. 4.2 of the report), where IR-limb still allows measurements in the mid-troposphere, where MW-limb fails to measure.

See MLS lack of penetration section above.

4) To my mind the trend analysis section needs to be improved. Although MIPAS tackles to deliver good trend analyses in tropical UT, this is a shortened result. Have you tested if the biases are better, if the trend analyses are restricted to longitude regions where convection is less pronounced (e.g. mid-Pacific,

Atlantic regions, excluding the warm pool region, central Africa and South America, where deep convection and cloud-coverage is most pronounced). Local trends in the UT/MT in tropical and subtropical areas would be valuable in principle.

See trends section above.

The reviewer is correct in pointing out that trends restricted to longitude regions where convection is less pronounced will be less affected (or barely affected) by cloud screening biases. For example, the following figure shows the yield at 200hPa for MIPAS and MLS water vapor and O_3 as well as the trends computed in the mid-Pacific where the MIPAS yields are less affected. As can be seen, the impact of the quality screening upon the trends in these region is minimal. In the paper we will add: Furthermore, MIPAS trend analysis can be restricted to regions less affected by deep convection (for example, the mid tropical Pacific) to minimize the quality screening effects.

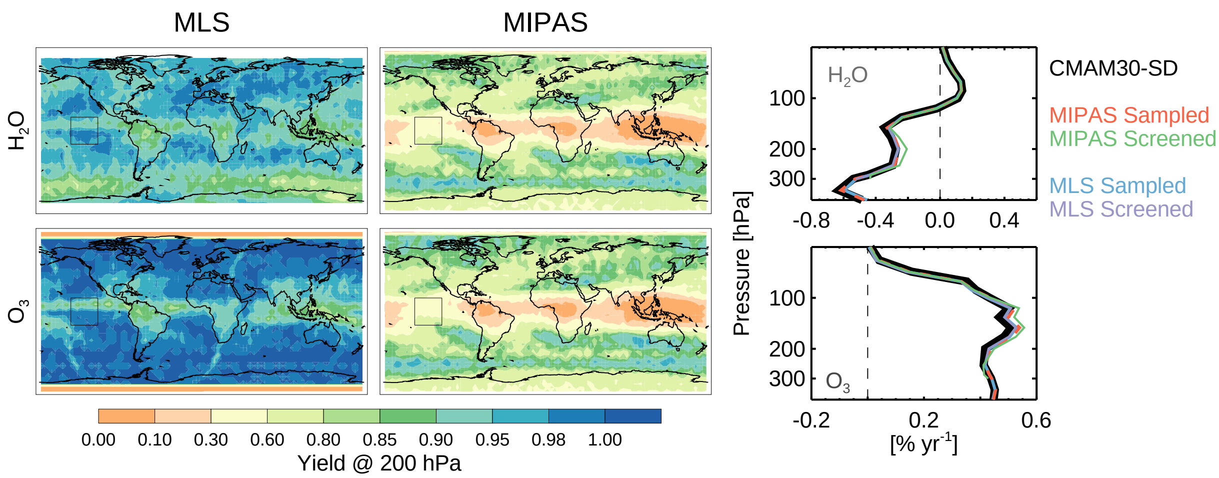


Figure: (right) Yield maps, (left) trends computed in the mid-pacific area shown in the maps for O_3 and H_2O.

5) I am also wondering why the authors argue that most of the biases are not caused by a lack of observations in convective (cloudy) areas in the tropics but are caused by a reduction in the overall sampling number. I think this is linked with each other, especially for water vapour with its extreme vertical and horizontal gradients. This objective would need further investigations for a final publication and is not sufficiently discussed by the only two sentences in section 3 (p5, line 25).

The sentence will be expanded as follows:
Although this resembles the expected dry bias in clear-sky tropospheric infrared measurements (e.g., Sohn et al., 2006; Yue et al., 2013) — that is, the fact that infrared instruments cannot measure cloudy regions where H2O is high, resulting in a dry bias — **the biases shown here are due to a combination of two factors: (1) high H2O values associated with deep convection (the screened-out locations might not necessarily be cloudy in the model fields however they are for the most part in the tropics, in regions of high H2O values, see Figure 1 for an example) and (2) due to the reduced sampling frequencies.** Note that this also applicable to other parameters …

6) Further, the quantification of the number of years needed to compensate for the biases introduced by the intensive cloud-clearing for IR limb sounder, which is only briefly mentioned in the Summary and Conclusion section, would be a very interesting topic for more detailed analyses and would improve the quality of the manuscript.

See trends section above.

Minor comments:

1) P2, line 23: Please, add a reference. We will add: [e.g., Livesey and Read (2000), Carlotti et al. (2001, 2006), Kiefer et al. (2010), Castelli et al. (2016)].]

References:
Livesey and Read (2000) - 10.1029/1999GL010964
Carlotti et al (2001) - 10.1364/AO.40.001872
Carlotti et al (2006) - 10.1364/AO.45.000716
Kiefer et al (2010) - 10.5194/amt-3-1487-2010
Castelli et al (2016) - 10.5194/amt-9-5499-2016

2) P3, line 3: Have you introduced CMAM30-SD already? We will change it to: CMAM30-SD (the Canadian Middle Atmosphere Model run in Specified Dynamics mode) is a coupled chemistry climate model nudged to the winds and temperatures of the ERA-Interim reanalysis…

3) P4, line 7: Please, explain in more detail, why the horizontal gradients are that important and why a different retrieval scheme might fail.
The following text will be added: Horizontal gradients are important because the line of sight extends around a thousand kilometers, crossing situations where the assumption of horizontal homogeneity is unrealistic (i.e., 1D retrievals). Several studies have discussed the advantages of including inhomogeneities along the line of sight for atmospheric retrievals [e.g., Livesey and Read (2000), Carlotti et al. (2001, 2006), Kiefer et al. (2010), Castelli et al. (2016)].

References:
Livesey and Read (2000) - 10.1029/1999GL010964
Carlotti et al (2001) - 10.1364/AO.40.001872
Carlotti et al (2006) - 10.1364/AO.45.000716
Kiefer et al (2010) - 10.5194/amt-3-1487-2010
Castelli et al (2016) - 10.5194/amt-9-5499-2016

4) P4, line 12: Are the ice cloud properties retrieved by a tomographic approach as well? Is a 2D or 3D tomographic approach, this should be mentioned.
The ice retrieval is not tomographic but rather 1D. The large degree of spatial inhomogeneity in cloud amount and cloud properties make tomographic retrieval approaches unstable in many situations, necessitating a simpler "1D" cloud retrieve whereby each limb view is considered to represent an independent cloud scene. The sentence will be changed to: These radiances are inverted using a 2D tomographic optimal estimation algorithm (Livesey et al., 2006) that allows the retrieval of temperature and composition.

5) This paper highlights technical aspects of the two sensors, so it might be helpful to include a minimum of instrumental details like field of view (FOV), vertical sampling (step size), retrieval resolution, and horizontal sampling. Otherwise the reader cannot judge on statements like ‘the vertical resolution is good enough to resolve these model fields’ without any reference.

We decided not to include details about FOV, vertical sampling, etc, so as not to confuse the reader with too many details about the measurements when this paper is actually using interpolated fields. Note that the vertical resolution of the model is in the original manuscript (P3 L15). The statement the reviewer refers to was changed to: Further, we used the vertical grid of the CMAM30-SD fields; that is, the impact of the vertical resolution of the measurements is not taken into account. However, note that, in this case, both instruments have similar vertical resolutions in the UTLS varying overall from 3 to 4 km.

6) Figure 1: please give some details why MIPAS fails to retrieve the most northern profiles. This looks like an artefact of the data processing. I would not expect strong horizontal gradients or clouds at these locations, like argued in the manuscript. Why fails MLS retrieval at some high northern latitudes as well? You should explain in more detail in the manuscript why and where retrieval my fail for both instruments.

An investigation of why both retrievals fail in those areas is outside the scope of this study.

7) Figure 3: Any explanation for the cold bias in the area of the south-pole for MIPAS? This is a bit confusing, because especially in the cold areas PSC are frequent and hamper proper retrievals for this part of the season, consequently I would expect a warm bias?! That would be assuming that CMAM30SD properly behaves in those regions, which may not be the case, further, this might be just a random effect due to sparse sampling after screening out the measurements.

Technical comments:

1) Figure 2, 3, and 4: A superimposed mean tropopause location will substantially improve the information content of the figures around the tropopause.

We will superimpose a mean 2005 thermal tropopause derived from MERRA2. See updated figure 2 above for an example.

2) Figure 2: please, give a few more details in the figure caption. E.g. Za-Zr or Zr-Za? We will add in brackets: N_{QS} / N_A where N_A is the number of measurements available and N_{QS} is the number of measurements left after applying the quality screening criteria.

3) Figure 5: The numbers on the x-axis are not well readable and should be sparser and separated. We changed it to every two years and increased the font size.

4) Figure 6+7: Temperature biases should be presented with a scaling of T (e.g. K x 10). Then the significant temperature bias in the tropics will become more obvious. In addition, a tropopause location might help here as well. The temperature biases will be shown as Kx10. We will change the x-label to [% or K*10] and the average climatological tropopause value for that latitude bin will be shown as a dashed black line.

The caption of figure 6 will be changed to: Vertical profiles of the coefficient of determination, the bias, and the linear fit slope for different latitude bands for the MIPAS vs CMAM30-SD scatter using the full satellite-sampled fields. Note that for clarity the temperature bias is shown as K*10. Blue, green, red, purple, and orange lines represent H_2O, O_3, CO, HNO_3, and temperature metrics. The dashed black lines show the mean 2005 thermal tropopause derived from MERRA2 for the particular latitude bands.

References:

Millán, L. F., Livesey, N. J., Santee, M. L., Neu, J. L., Manney, G. L., and Fuller, R. A.: Case studies of the impact of orbital sampling on stratospheric trend detection and derivation of tropical vertical velocities: solar occultation vs. limb emission sounding, Atmos. Chem. Phys., 16, 11521-11534, https://doi.org/10.5194/acp-16-11521-2016, 2016.

Spang, R., Arndt, K., Dudhia, A., Höpfner, M., Hoffmann, L., Hurley, J., Grainger, R. G., Griessbach, S., Poulsen, C., Remedios, J. J., Riese, M., Sembhi, H., Siddans, R., Waterfall, A., and Zehner, C.: Fast cloud parameter retrievals of MIPAS/Envisat, Atmos. Chem. Phys., 12, 7135-7164, https://doi.org/10.5194/acp-12-7135-2012, 2012.